PLOS | GENETICS

RESEARCH ARTICLE

A New Method for Detecting Associations with Rare Copy-Number Variants

Jung-Ying Tzeng[1,2], Patrik K. E. Magnusson[3], Patrick F. Sullivan[3,4], The Swedish Schizophrenia Consortium[¶], Jin P. Szatkiewicz[3]*

1 Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **2** Department of Statistics, National Cheng-Kung University, Tainan, Taiwan, **3** Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America, **4** Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

¶ Membership in The Swedish Schizophrenia Consortium is listed in the Acknowledgments.
* jin_szatkiewicz@med.unc.edu

OPEN ACCESS

Citation: Tzeng J-Y, Magnusson PKE, Sullivan PF, The Swedish Schizophrenia Consortium, Szatkiewicz JP (2015) A New Method for Detecting Associations with Rare Copy-Number Variants. PLoS Genet 11 (10): e1005403. doi:10.1371/journal.pgen.1005403

Editor: Hakon Hakonarson, Children's Hospital of Philadelphia, UNITED STATES

Received: January 29, 2015

Accepted: June 30, 2015

Published: October 2, 2015

Data Availability Statement: Two types of data were used in this work and the data-sharing methods are: (1) Example simulated datasets are available at http://www4.stat.ncsu.edu/~tzeng/software.php. (2) The real data from the Swedish schizophrenia study is available through the NIMH repository (controlled access repository), which ensures that the data from the NIMH supported human subject studies can be shared in a timely manner to the wider scientific community, provided that the confidentiality and privacy of study participants are protected. Detailed information about the NIMH repository can be found at http://www.nimh.nih.gov/funding/clinical-trials-for-

Abstract

Copy number variants (CNVs) play an important role in the etiology of many diseases such as cancers and psychiatric disorders. Due to a modest marginal effect size or the rarity of the CNVs, collapsing rare CNVs together and collectively evaluating their effect serves as a key approach to evaluating the collective effect of rare CNVs on disease risk. While a plethora of powerful collapsing methods are available for sequence variants (e.g., SNPs) in association analysis, these methods cannot be directly applied to rare CNVs due to the CNV-specific challenges, i.e., the multi-faceted nature of CNV polymorphisms (e.g., CNVs vary in size, type, dosage, and details of gene disruption), and etiological heterogeneity (e.g., heterogeneous effects of duplications and deletions that occur within a locus or in different loci). Existing CNV collapsing analysis methods (a.k.a. the burden test) tend to have suboptimal performance due to the fact that these methods often ignore heterogeneity and evaluate only the marginal effects of a CNV feature. We introduce CCRET, a random effects test for collapsing rare CNVs when searching for disease associations. CCRET is applicable to variants measured on a multi-categorical scale, collectively modeling the effects of multiple CNV features, and is robust to etiological heterogeneity. Multiple confounders can be simultaneously corrected. To evaluate the performance of CCRET, we conducted extensive simulations and analyzed large-scale schizophrenia datasets. We show that CCRET has powerful and robust performance under multiple types of etiological heterogeneity, and has performance comparable to or better than existing methods when there is no heterogeneity.

Author Summary

Copy number variants (CNVs) are the gain or loss of DNA segments in the genome that can vary in dosage, length and details of gene disruptions. Rare CNVs have been shown to be associated with neuropsychiatric disorders both collectively and at specific loci. To evaluate the collective effects of rare CNVs on disease risk, sophisticated association methods

researchers/datasets/index.shtml Procedures to obtain data from the NIMH repository can be found at http://www.nimh.nih.gov/funding/clinical-trials-for-researchers/datasets/nimh-procedures-for-requesting-data-sets.shtml

Funding: For this work, JPS is supported by NIH grants (K01MH093517 and R21MH104831). JYT is supported by a NIH grant (P01CA142538). PFS is supported by a NIH grant (R01MH077139). The Swedish Twin Registry is financially supported by Karolinska Institutet. Funding support for the Swedish schizophrenia study was provided by NIH grants (R01MH077139 and R01 MH095034), the Stanley Center for Psychiatric Research, the Stanley Medical Research Institute, the Sylvan Herman Foundation, the Karolinska Institutet, Karolinska University Hospital, the Swedish Research Council, the Swedish County Council, the Söderström Königska Foundation, and the Netherlands Scientific Organization (NWO 645-000-003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: In the interests of full disclosure, PFS was on the SAB of Expression Analysis (Durham, NC, USA). The other authors report no conflicts.

are needed to pool information across CNV loci while handling CNV-specific properties; however, such methods are under-developed. To address these challenges, we have developed a new collapsing method for rare CNVs named CCRET. CCRET is a random effects approach applicable to variants measured on a multi-categorical scale, collectively modeling the effects of multiple CNV features, and is robust to etiological heterogeneity. Multiple confounders can be simultaneously corrected. To evaluate the performance of CCRET, we conducted extensive simulation and analyzed large-scale schizophrenia datasets. We demonstrate the robustness, validity and utility of CCRET under a variety of scenarios.

Introduction

Copy-number variants (CNVs), such as deletions and duplications, are changes in the number of DNA copies (in comparison to the reference) and are a major source of genetic variation in the human genome [1–3]. While a sequence variant (e.g., a SNP) affects a single nucleotide, a CNV affects a region $>$ 1 Kb. CNV may alter the "dosage" of one or more genes or regulatory regions in the deleted or duplicated region, which can consequently exert a profound effect on the risk for human disease. Genomic evaluation of CNVs has established a role for rare ($<$1%) CNVs in the etiology of psychiatric disorders, such as schizophrenia, bipolar, and autism [4–6]. Eight rare CNVs of strong effects are now established risk factors for psychiatric disorders (e.g., 16p11.2, 22q11.2, genotypic relative risk 4–20) [4, 5]. However, the bulk of CNVs' contribution to disease risk remains unknown due to a modest effect size or the rarity of the CNV. As psychiatric disorders are polygenic, collapsing methods [7], which collapse multiple variants into a group and evaluate their collective effect on disease risk, serve as key approaches to the analysis of rare CNVs [4, 5]. By accumulating information across multiple rare variants (e.g., counting the number of mutations for each individual), collapsing methods can have an enhanced power to detect genetic variants that are hard to detect individually but collectively show a significant impact. For example, multiple studies have confirmed a greater genomewide burden of rare CNVs in schizophrenia cases compared with controls [8–12]. Enrichment analyses of genes impacted by rare CNVs implicated several biological pathways important to schizophrenia, including those previously associated with schizophrenia through common variation and exome sequencing (e.g., calcium channel signaling and binding partners of the fragile X mental retardation protein) [12].

For rare sequence variants such as SNPs, a plethora of powerful methods are available to perform collapsing analysis. Depending on the approaches used to model genetic effects and the procedures used to collapse the information across loci, these methods can be classified into two major categories: fixed effects methods (e.g., CMC [13], VT [14]) and random effects methods (e.g., C-alpha [15], SKAT [16], SimReg [17, 18]). A detailed review can be found in Pongpanich et al. [19] and Lee et al. [20]. Briefly, fixed effects methods collapse information at genotype level and assess the mean level of the genetic effects via fixed effects modeling; it is the optimal approach if the effects of different loci are additive and of a similar size. In contrast, random effects methods collapse information at similarity level and assess the variance level of the genetic effects via random effects modeling. Random effects methods are more powerful than fixed effects methods when the variants have different effects (e.g., mixture of positive, negative and neutral effects).

However, SNP collapsing methods cannot be straightforwardly applied to CNVs due to several CNV-specific challenges. First, copy number is measured on a multi-categorical scale (e.g., duplication, normal copy and deletion) while SNP collapsing methods assume binary events (e.g., mutation vs. no mutation). Second, CNV can vary in dosage (i.e., the copy number of a CNV), length (i.e., the segment size of a CNV) and details of gene disruptions (i.e., the number of genes that a CNV intersects), and each of these "features" affects CNVs' impact on disease risk. For example, in schizophrenia, deletions were enriched in cases to a greater extent than duplications, and the largest CNVs ($>$ 500 kb) were enriched in cases to a greater extent than other size categories [8–12]. Girirajan et al. [21] found that the total duplication length is significantly elevated in autism cases compared with controls. On the other hand, SNP collapsing methods target only one feature (i.e., mutation burden). Third, etiological heterogeneity is often observed in CNVs. While SNPs only exhibit between-locus heterogeneity, the etiological heterogeneity of CNVs can occur both between loci (e.g., CNVs across different loci have different effects on disease risk) and within a locus (i.e. different dosages of CNV alleles within the same locus have different effects). (Please see "Input data format" in the Method Section for detailed definitions of "locus".) For example, the 22q11.2 deletion is a known risk factor for schizophrenia [4, 5, 22–24], whereas the reciprocal 22q11.2 duplication is potentially a protective factor [25]. Microduplications of gene *VIPR2* increase the risk for schizophrenia, where both tandem duplication (copy number 3) and triplication (copy number 4) were observed with triplication potentially conveying higher risk than duplication [26]. The etiological heterogeneity may occur more frequently as CNV detection technologies continue to improve, allowing the accurate detection of small CNVs. Naïve collapsing of a mixture of neutral, risk, and protective variants between loci or within a locus can cancel signals and lead to power loss. Random effects SNP-collapsing methods have the potential to address between-locus etiological heterogeneity in CNV analyses; however, because these methods record the genetic information using the number of mutant events, they cannot deal with the within-locus etiological heterogeneity observed in multi-categorical scale CNVs.

Collapsing methods based on fixed effects approaches have been developed for rare CNVs [9, 27]. Specifically, burden-style tests [9] examine CNV events to evaluate whether an increased rate or the size of the CNVs increases disease risk. Rare CNVs are typically aggregated based on a certain event of interest and then summarized by the event counts, such as the number of deletions (copy number <2), the number of duplications (copy number >2), the number of CNVs (copy number$\neq 2$), or the number of genes intersected by the CNVs. The association is evaluated by comparing the event rates between cases and controls. However, this approach falls short in three aspects. First, the researcher must pre-select a CNV event (e.g., deletion or duplication or both) and summarize the data using the appropriate event counts. Second, it ignores etiological heterogeneity. Finally, it only evaluates the marginal effects of a CNV feature, which may or may not be a valid approach depending on the collapsing unit. For example, while the gene-count burden test is valid for a genomewide collapsing, it can result in spurious association if based on a gene set [27]. To address the last issue, enrichment-style tests [27] have been proposed to assess the conditional effect of a CNV feature; such a test compares the counts of genes within a pre-specified gene set intersected by CNVs in cases with the counts in controls while controlling for case-control differences in genomewide CNV rate and size. Although it uses a joint modeling to avoid spurious association, this method does not address the issue of etiological heterogeneity and still needs to summarize CNV data into counts based on the feature of interest.

We have developed a new collapsing method for the analysis of rare CNVs that is applicable to variants measured on a multi-categorical scale, collectively modeling the effects of multiple CNV features, and is robust to etiological heterogeneity. Our method is called CCRET (CNV Collapsing Random Effects Test, pronounced as "se-cret"). Specifically, we use random effects to model the CNV effect of interest, which, for SNP collapsing analysis, has been shown to be more robust than a fixed effects approach when the complexity of genetic architecture increases [19, 20]. When aggregating information through CNV similarity, we define "locus" units for each CNV feature to retain the "locus"-specific effect during collapsing, and use factorized design vectors for multi-categorical CNV features to quantify similarity without dichotomizing the data as is done in the fixed effects methods. CCRET can simultaneously correct for covariates such as batch effects and population substructures. CCRET can be viewed as an extension of the SNP-set random effects methods (e.g., C-alpha [15], SKAT [16], SimReg [17, 18]) applicable to CNV data. To evaluate the performance of CCRET, we conducted extensive simulations under various scenarios and analyzed large-scale CNV data from the Swedish schizophrenia study. Our results suggest that, compared with the existing (fixed effects) CNV collapsing methods implemented in PLINK [9, 27, 28], our random effects approach has a stable, powerful and robust performance under multiple types of etiological heterogeneity, and has a comparable or better performance when there is no heterogeneity.

Results

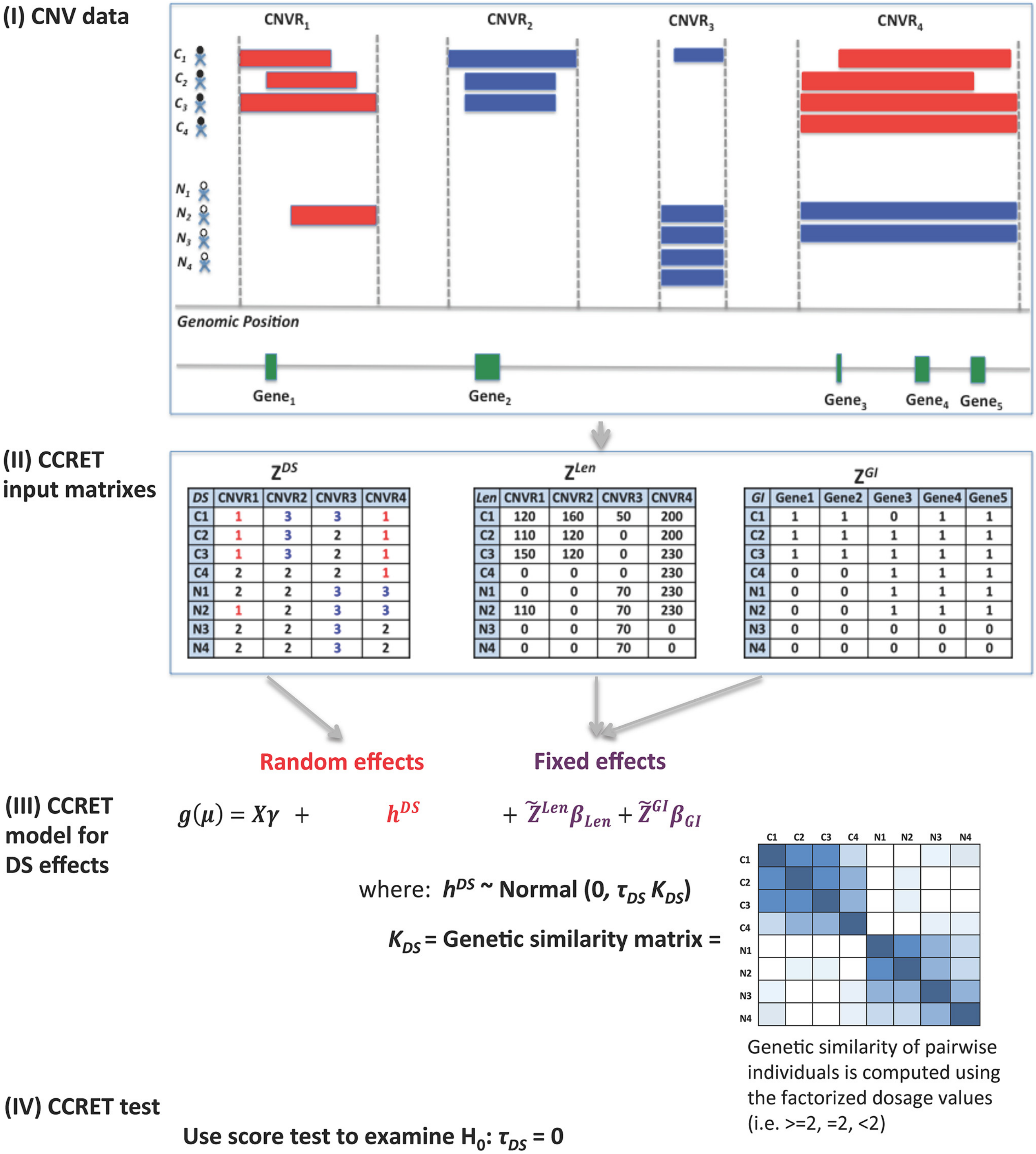
Overview of CCRET (CNV Collapsing Random Effects Test)

Fig 1 provides an overview of the CCRET method using the dosage effects model as an example. CCRET aims to detect any association of the aggregated CNV effect with disease risk and has the following key features. First, CCRET converts the source CNV data to three input matrixes in order to store the different features of CNVs, i.e., dosage ("DS"), length ("Len"), and gene intersection ("GI"). For "DS" and "Len" matrixes, we use CNV regions as the "locus" unit. For "GI" matrix, we use genes as the "locus" units. Second, CCRET models the covariates and background CNV features using fixed effects as did in Raychaudhuri et al [27], and models the CNV feature of interest using random effects in order to retain the locus-specific details and to account for both between-locus and within-locus etiological heterogeneity. Third, CCRET quantifies the genetic similarity between any two individuals based on the CNV feature of interest, which is then used to depict the covariance among the CNV effects of different individuals (i.e, the more similar the genetic feature between two individuals is, the more correlated their CNV effects would be). When calculating genetic similarity, we factorize the multi-categorical allele values recorded in the input matrices. Consequently, alleles with opposite effects within a locus are not lumped together when computing similarity, which makes CCRET robust against within-locus heterogeneity. In contrast, SNP-collapsing random effects methods (e.g., C-alpha [15], SKAT [16], SimReg [17, 18]) do not address within-locus heterogeneity. Finally, under the mixed effects model framework, the aggregate CNV effect can be evaluated by examining the significance of the variance component. In contrast, fixed effects methods test the aggregated CNV effect by examining if the common (e.g., averaged) effect is equal to zero.

Fig 2 provides an overview of the evaluative analyses carried out in this work. We conducted two sets of simulations under a variety of scenarios and conducted real data analysis using large-scale schizophrenia datasets. We evaluated the performance of CCRET in comparison to the fixed effects CNV-collapsing methods implemented in PLINK [9, 27, 28].

Simulation studies for performance evaluations

We conducted two sets of simulations to evaluate the performance of our method under a variety of scenarios: (I) causal dosage effects and (II) causal gene intersection (GI) effects. In each simulation, the genotypes of rare CNVs were taken from the TwinGene study [29, 30], which used a cross-sectional sampling design and selected 6,611 unrelated subjects born between 1911 and 1958 from the Swedish Twin Registry [31–33] (STR) for genomic characterization. These samples included one member per monozygotic twin pair and a randomly selected



Z^{DS}

DS	CNVR1	CNVR2	CNVR3	CNVR4
C1	1	3	3	1
C2	1	3	2	1
C3	1	3	2	1
C4	2	2	2	1
N1	2	2	3	3
N2	1	2	3	3
N3	2	2	3	2
N4	2	2	3	2

Z^{Len}

Len	CNVR1	CNVR2	CNVR3	CNVR4
C1	120	160	50	200
C2	110	120	0	200
C3	150	120	0	230
C4	0	0	0	230
N1	0	0	70	230
N2	110	0	70	230
N3	0	0	70	0
N4	0	0	70	0

Z^{GI}

GI	Gene1	Gene2	Gene3	Gene4	Gene5
C1	1	1	0	1	1
C2	1	1	1	1	1
C3	1	1	1	1	1
C4	0	0	1	1	1
N1	0	0	1	1	1
N2	0	0	1	1	1
N3	0	0	0	0	0
N4	0	0	0	0	0

$$g(\mu) = X\gamma + h^{DS} + \tilde{Z}^{Len}\beta_{Len} + \tilde{Z}^{GI}\beta_{GI}$$

where: $h^{DS} \sim \text{Normal}(0, \tau_{DS} K_{DS})$

K_{DS} = Genetic similarity matrix =

Genetic similarity of pairwise individuals is computed using the factorized dosage values (i.e. >=2, =2, <2)

Use score test to examine H_0: $\tau_{DS} = 0$

Fig 1. An overview of the CCRET method. The dosage effects model was used as an example. $C_{1\text{-}4}$: cases, $N_{1\text{-}4}$: controls, CNVR: copy number variation region, red rectangle: deletion, blue rectangle: duplication, green rectangle: gene. DS: dosage, Len: length, GI: gene intersection.

doi:10.1371/journal.pgen.1005403.g001

member per dizygotic twin pair. DNA was extracted from peripheral venous blood for all subjects. The samples were genotyped on the Illumina OmniExpress beadchip for 72,881 SNP markers using standard protocol; CNV calling was performed using PennCNV [34] (version

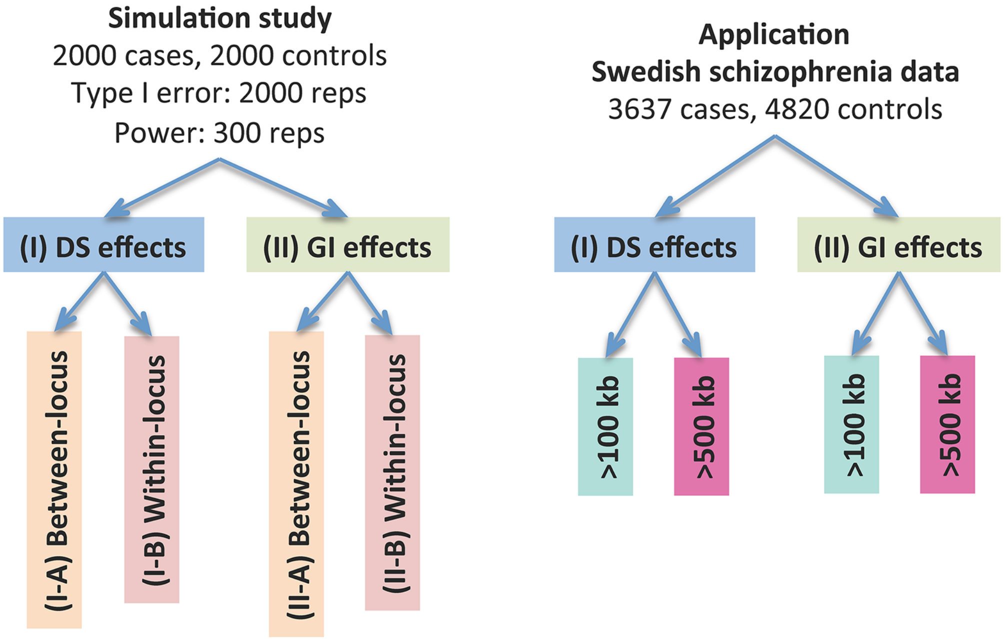


Fig 2. Performance evaluation of the CCRET method. "Between-locus": between-locus heterogeneity; "Within-locus": within-locus heterogeneity.

doi:10.1371/journal.pgen.1005403.g002

June 2011) with recommended model parameters. We randomly selected 2,000 subjects from the TwinGene study and formed 1,757 CNV regions (CNVRs) genome-wide with frequency $<1\%$. The detailed data processing procedure is described in *S1 Text*. In each simulation setting, we used case-control sampling to collect 2,000 cases and 2,000 controls, evaluated the type I error rates based on 2,000 replications and evaluated the power based on 300 replications.

Simulation I: evaluating dosage effects. We considered causal dosage effects with two types of etiological heterogeneity between deletions and duplications: (A) between-locus heterogeneity, and (B) within-locus heterogeneity. We simulated the binary phenotype of individual i from the Bernoulli distribution with π_i as the probability of being a case and $logit(\pi_i) = -2 + \sum_m \beta_m^{DS.Dup} \times z_{im}^{DS.Dup} + \sum_m \beta_m^{DS.Del} \times z_{im}^{DS.Del}$, where $z_{im}^{DS.Dup} = 1$ if individual i at causal locus m has a duplication and 0 otherwise; $z_{im}^{DS.Del}$ is defined in a similar fashion for deletion events; $\beta_m^{DS.Dup}$ and $\beta_m^{DS.Del}$ are the log of the odds ratios (ORs) of causal locus m for duplications and deletions, respectively; $\beta_m^{DS.Dup}$ and $\beta_m^{DS.Del}$ shared the same absolute values for all causal loci but were positive if associated with increased disease risk and negative if protective effects. We compared CCRET to the collapsing methods implemented in PLINK. Specifically, PLINK command "—cnv-indiv-perm" was applied to duplications only (referred to as PLINK.dup), deletions only (referred to as PLINK.del), and deletions and duplications combined (referred to as PLINK.all). The PLINK p-values were computed using 10,000 permutations.

Simulation I-A: between-locus heterogeneity of the dosage simulation. Among the 1,757 CNVRs, there are 766 loci with duplication genotypes only (dosage 3 or 4, "DupOnly" hereafter), 840 loci with deletion genotypes only (dosage 0 or 1, "DelOnly" hereafter), and 151 loci with both deletion and duplication genotypes (dosage 0,1,3,4, "DupDel" hereafter). We considered 600 causal loci where 300 loci were randomly selected from the 766 DupOnly loci and 300 loci were randomly from the 840 DelOnly loci. We considered 6 heterogeneity models with different proportions of risk-associated ($\beta_m^{DS.\bullet} > 0$) or protective ($\beta_m^{DS.\bullet} < 0$) effects among

Table 1. Type I error rates for evaluating dosage effects (nominal alpha = 0.05).

Model	CCRET	PLINK.all	PLINK.dup	PLINK.del
(A) Between-locus heterogeneity	0.035	0.046*	0.057	0.041
		(0.047)**	(0.053)	(0.055)
(B) Within-locus heterogeneity	0.041	0.051	0.057	0.043

*: Type I error rates based on PLINK 2-sided tests

**: Type I error rates in parentheses are based on PLINK 1-sided tests.

doi:10.1371/journal.pgen.1005403.t001

causal loci. The first three served as the baseline models which favor the PLINK burden tests, i.e., (1) all DupOnly risk-associated and all DelOnly protective (i.e., $\beta^{DS.Dup}$ +/− = 100/0 and $\beta^{DS.Del}$ +/− = 0/100); (2) all DupOnly protective and all DelOnly risk-associated (i.e., $\beta^{DS.Dup}$ +/− = 0/100 and $\beta^{DS.Del}$ +/− = 100/0); (3) all causal loci risk-associated (i.e., no heterogeneity) with $\beta^{DS.Dup}$ +/− = 100/0 and $\beta^{DS.Del}$ +/− = 100/0. The remaining three models consider different levels of heterogeneity, i.e., (4) $\beta^{DS.Dup}$ +/− = 70/30 and $\beta^{DS.Del}$ +/− = 30/70; (5) $\beta^{DS.Dup}$ +/− = 30/70 and $\beta^{DS.Del}$ +/− = 70/30, and (6) $\beta^{DS.Dup}$ +/− = 50/50 and $\beta^{DS.Del}$ +/− = 50/50. For each scenario, we assumed a constant OR for casual loci and considered OR ranging between 1 and 7, chosen based on the empirical evidence of pathogenic CNVs in psychiatric disorders [4, 5].

The type I error rates were around the nominal level for all methods under between-locus, although the results of CCRET are slightly conservative (Table 1). For power analyses, we first compared CCRET to PLINK 1-sided tests, which assess whether the event rate is higher in cases than in controls (Fig 3). In models (1) to (3) (Fig 3 upper panel, where duplications (deletions) have the same effects and PLINK tests would be the most powerful), CCRET provided comparable power to the best PLINK methods while the best PLINK method varied. Specifically, in (1) where all duplications were risk-associated and all deletions were protective, PLINK.dup was the best method as expected; similarly, in (2), PLINK.del was the best method as expected; and in (3), PLINK.all was the best method, because PLINK.all used all available information whereas PLINK.dup or PLINK.del only used a subset of the total events. In Models (4) to (6), (Fig 3 lower panel, where different combinations existed of risk-associated and protective effects in β^{Dup} and β^{Del}), CCRET consistently yielded the best power or yielded power comparable to the best PLINK method. From Fig 3, we also see that the performance of PLINK 1-sided tests was highly dependent on the underlying effect mechanisms. Specifically, the best PLINK 1-sided tests were those focused on the CNV events with risk-associated effect, and those PLINK 1-sided tests that focused on the events with protective effects had no power. In the presence of etiological heterogeneity (i.e., Models (1), (2), and (4) through (6)), the performance of PLINK.all was hard to predict; roughly speaking, it tended to be in and between PLINK.dup and PLINK.del, and the power somewhat depended on the relative proportion of the causal risk-associated and causal protective CNVRs.

In Fig 4, we compare CCRET to PLINK 2-sided tests. For PLINK 2-sided tests, we observed a pattern of the relative performance of PLINK and CCRET similar to that seen in Fig 3. The only exception was that PLINK.dup and PLINK.del had good power under heterogeneity models (1) and (2). In sum, PLINK 2-sided tests are more robust than PLINK 1-sided tests when the underlying effect patterns are unknown; and therefore we present the comparisons of CCRET and PLINK 2-sided tests for the remaining simulation studies.

Simulation I-B: within-locus heterogeneity of the dosage simulation. We selected 38 loci out of the 151 DupDel loci to be causal and the selected causal loci tended to have a similar number of duplications and deletions. We considered 5 heterogeneity models, where deletions and duplications had different combinations of risk-associated (R), protective (P) or neutral

(N) effects on the phenotypes. Specifically, the five effect combinations considered were (Dup, Del) = (R,N), (N,R), (P,R), (R,P) and (R,R) (i.e., no heterogeneity). For example, "(Dup, Del) = (P,R)" indicates that, within each of the 38 loci, duplications had protective effects, whereas deletions had risk-associated effects. For each model, we set a constant OR for the casual loci and considered OR ranging between 1 and 20, chosen based on empirical evidences [4, 5].

Table 1 shows that the type I error rates were around the nominal level for all methods. The power results (Fig 5) show that CCRET had power comparable to or better than the best PLINK test across all heterogeneity models considered. Again, the best PLINK test varied across heterogeneity models, but overall it focused on the CNV allele with risk-associated effects. The PLINK tests that focused on the protective (neutral) alleles had low (no) power. PLINK.all had power similar to and between PLINK.dup and PLINK.del, except in the case where within-locus heterogeneity did not exist, i.e., (R,R).

Simulation II: evaluating GI effects. In Simulation II, we considered causal GI with heterogeneous effects between a duplication GI and deletion GI. In the TwinGene data, we focused on 668 protein-coding genes in the neuronal postsynaptic density (PSD) because previous researchers have reported enrichment of rare CNVs in the PSD genes in schizophrenia cases [12, 35]. CNVs intersected 69 of the 668 genes, where 35 genes were intersected by duplications

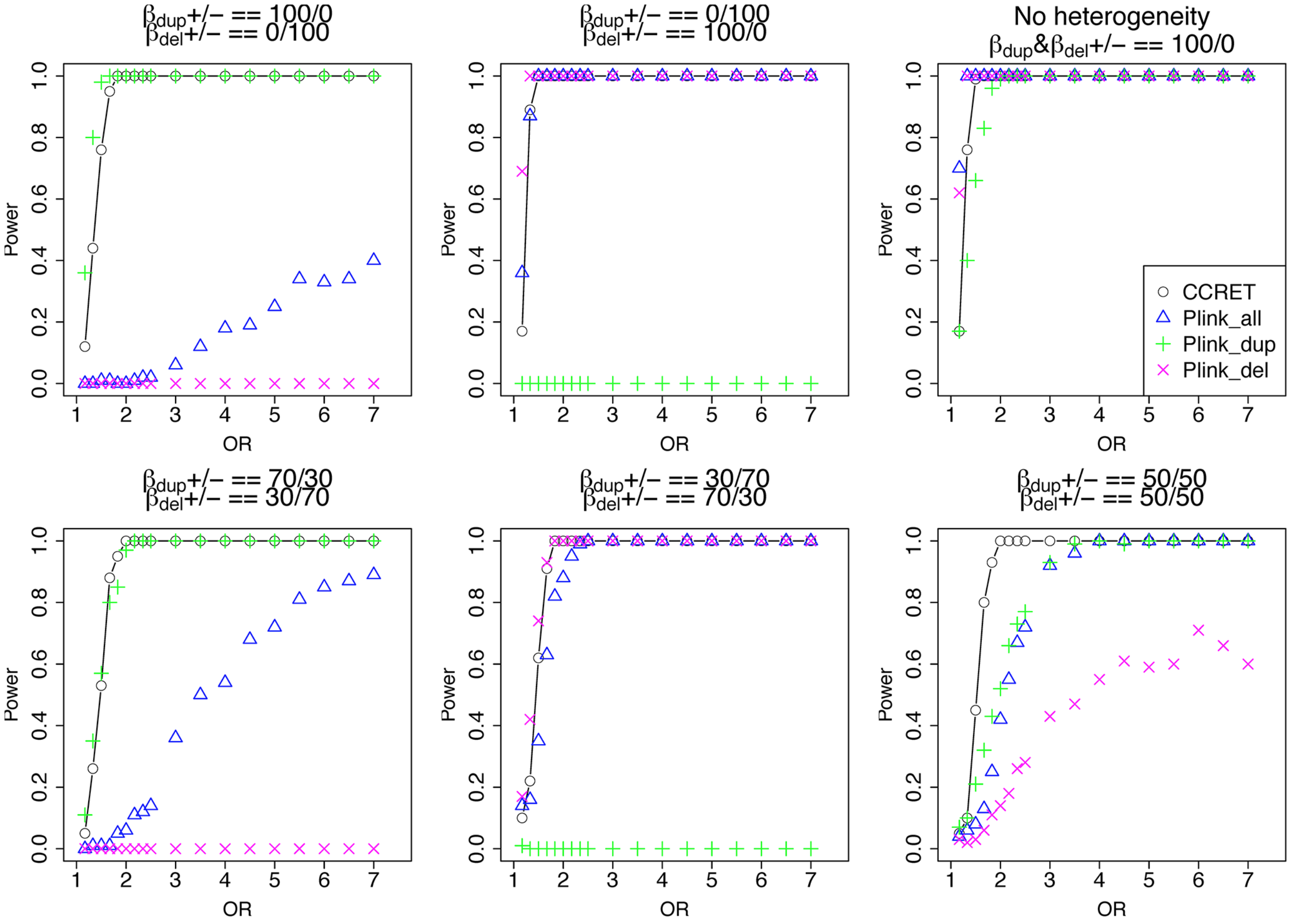


Fig 3. Power comparison between CCRET and PLINK 1-sided tests for simulation I-A: between-locus heterogeneity of the dosage simulation. As detailed in section "Simulation Design", we considered 6 heterogeneity models with different proportions of risk-associated ($\beta_m^{DS\cdot} > 0$) or protective ($\beta_m^{DS\cdot} < 0$) effects among causal loci. Black **O** line: CCRET; Blue **Δ**: PLINK 1-sided test analyzing deletions and duplications combined; Green **+**: PLINK 1-sided test analyzing only duplications; Magenta **x**: PLINK 1-sided test analyzing only deletions.

doi:10.1371/journal.pgen.1005403.g003

only ("DupOnly genes" hereafter), 26 genes were intersected by deletions only ("DelOnly genes" hereafter), and 8 genes were intersected by both deletions and duplications ("DupDel genes" hereafter). Similar to the first simulation study, we considered (A) between-gene heterogeneity and (B) within-gene heterogeneity. We simulated the binary phenotype of individual i from the Bernoulli distribution with π_i as the probability of being a case and $logit(\pi_i) = -2.5 + \sum_m \beta_m^{GI.Dup} \times z_{im}^{GI.Dup} + \sum_m \beta_m^{GI.Del} \times z_{im}^{GI.Del} + \beta^{CNV} \times z_{CNV,i} + \beta^{Len} \times z_{Len,i}$, where $z_{im}^{GI.Dup} = 1$ if causal gene m is intersected by a duplication and 0 otherwise; $z_{im}^{GI.Del}$ is defined similarly for deletion; $z_{CNV,i}$ is the total number of CNV events; and, $z_{Len,i}$ is the mean size of the CNVs measured in kb. The regression coefficients $\beta_m^{GI.Dup}$ and $\beta_m^{GI.Del}$ are the log (OR)'s of causal gene m for a duplication intersection and a deletion intersection, respectively; $\beta_m^{GI.Dup}$ and $\beta_m^{GI.Del}$ shared the same absolute values for all causal loci but were positive if risk-associated effects and negative if protective effects; β^{CNV} and β^{Len} were set to be log(1.5). We compared CCRET with the PLINK enrichment test as described in Raychaudhuri et al. [27]. We implemented the procedure in R and reported the 2-sided asymptotic p-values as done in the default option of PLINK "—cnv-enrichment-test". We performed the enrichment analysis for deletions and duplications combined (referred to as PLINK.enrich), deletions only (PLINK.enrich.del), and duplications only (PLINK.enrich.dup).

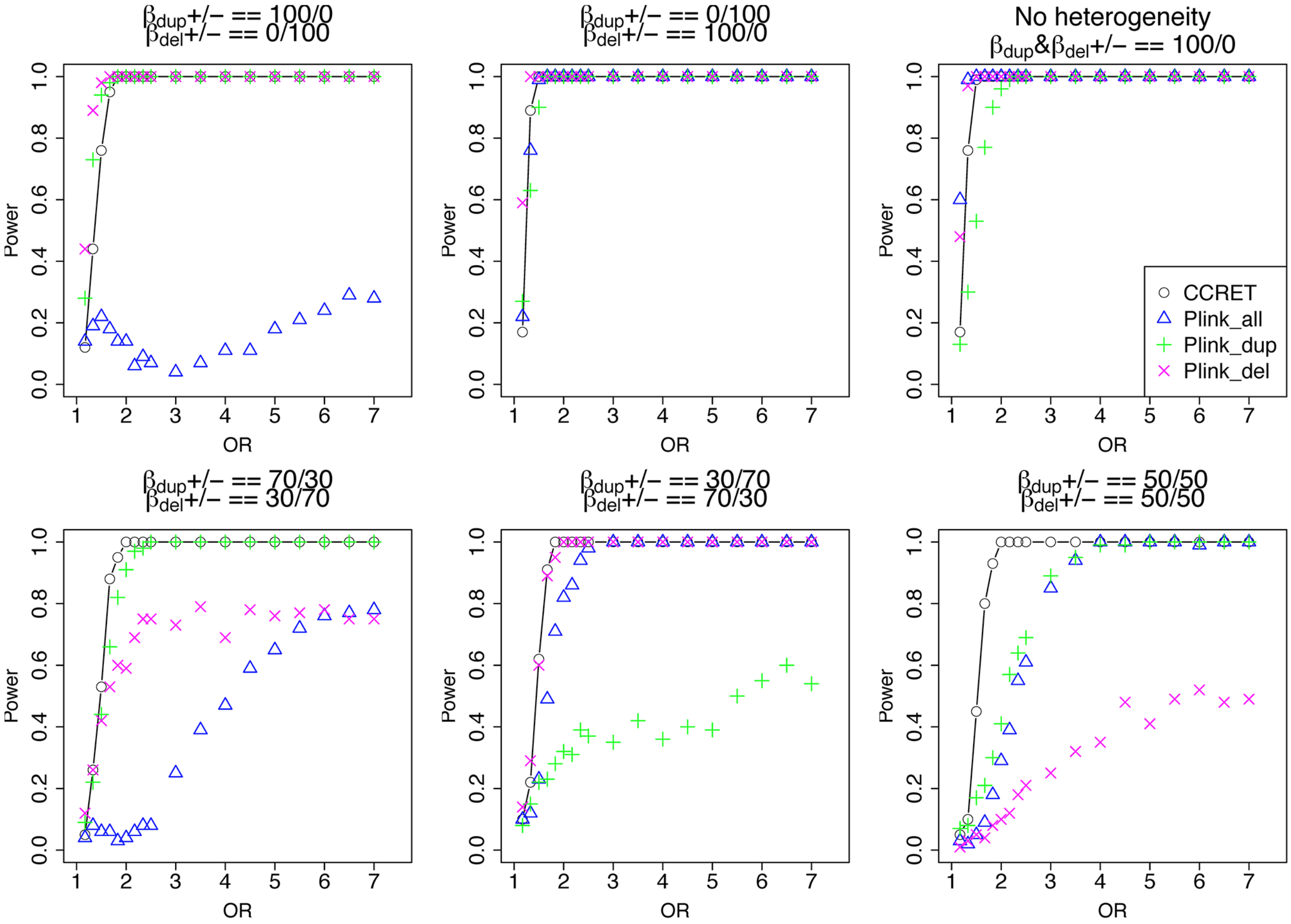


Fig 4. Power comparison between CCRET and PLINK 2-sided tests for simulation I-A: between-locus heterogeneity of the dosage simulation, under 6 heterogeneity models as in Fig 3. Black **O** line: CCRET; Blue **Δ**: PLINK 2-sided test analyzing deletions and duplications combined; Green **+**: PLINK 2-sided test analyzing only duplications; Magenta **x**: PLINK 2-sided test analyzing only deletions.

doi:10.1371/journal.pgen.1005403.g004

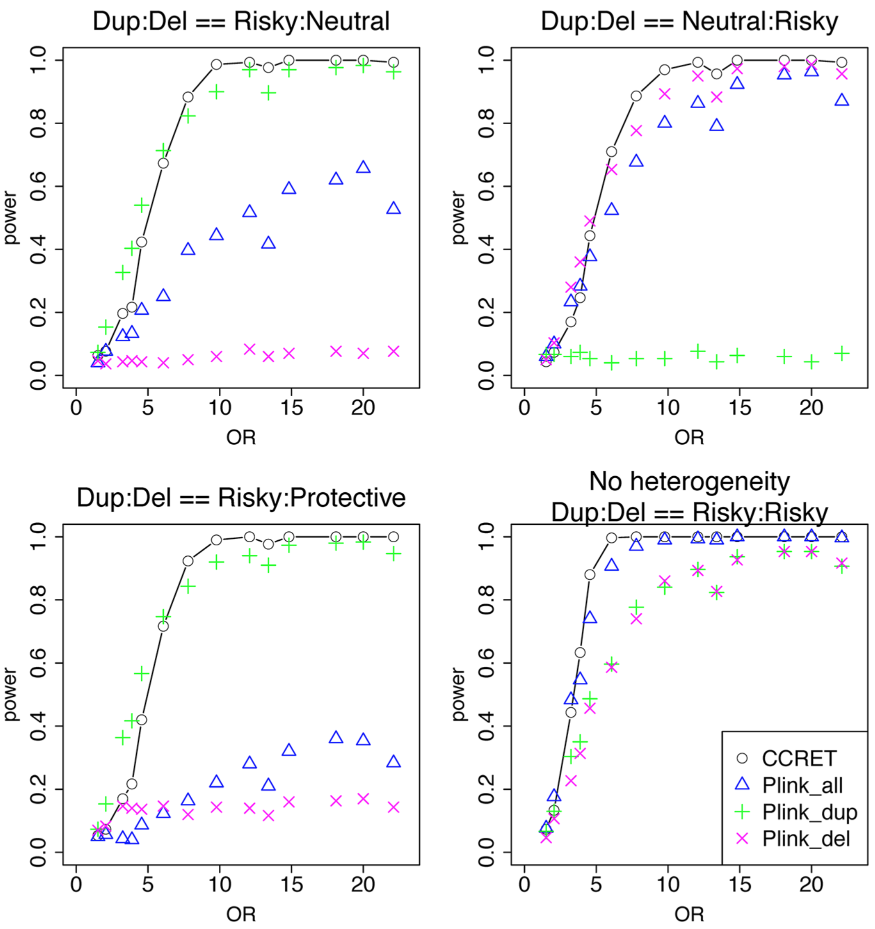


Fig 5. Power comparison between CCRET and PLINK 2-sided tests for simulation I-B: within-locus heterogeneity of the dosage simulation, under 5 heterogeneity models as detailed in section "Simulation Design". Black **O** line: CCRET; Blue **Δ**: PLINK 2-sided test analyzing deletions and duplications combined; Green **+**: PLINK 2-sided test analyzing only duplications; Magenta **x**: PLINK 2-sided test analyzing only deletions.

doi:10.1371/journal.pgen.1005403.g005

Simulation II-A: between-locus heterogeneity of the GI simulation. We randomly selected 26 genes from the DupOnly genes and 26 genes from the DelOnly genes, and used the 52 genes as causal loci. We considered 6 heterogeneity models similar to the models considered in the dosage simulation: (1) DupOnly risk-associated and DelOnly protective (i.e., $\beta^{GI.Dup}$ +/− = 100/0 and $\beta^{GI.Del}$ +/− = 0/100); (2) DupOnly protective and DelOnly risk-associated (i.e., $\beta^{GI.Dup}$ +/− = 0/100 and $\beta^{GI.Del}$ +/− = 100/0); (3) all causal loci risk-associated (i.e., no heterogeneity with $\beta^{GI.Dup}$ +/− = 100/0 and $\beta^{GI.Del}$ +/− = 100/0); (4) $\beta^{GI.Dup}$ +/− = 70/30 and $\beta^{GI.Del}$ +/− = 30/70; (5) $\beta^{GI.Dup}$ +/− = 30/70 and $\beta^{GI.Del}$ +/− = 70/30; and (6) $\beta^{GI.Dup}$ +/− = 50/50 and $\beta^{GI.Del}$ +/− = 50/50. We set a constant effect size for all casual loci and considered the OR ranging between 1 and 20, chosen based on empirical evidences [4, 5].

Table 2 shows that the type I error rates were around the nominal level for PLINK.enrich and CCRET, although CCRET was slightly conservative. Fig 6 suggests a consistent power gain

Table 2. Type I error rates for evaluating GI effects (nominal alpha = 0.05).

Model	CCRET	PLINKenrich	PLINK enrich.dup	PLINK enrich.del
(A) Between-locus heterogeneity	0.041	0.044	0.051	0.042
(B) Within-locus heterogeneity	0.043	0.051	0.052	0.043

doi:10.1371/journal.pgen.1005403.t002

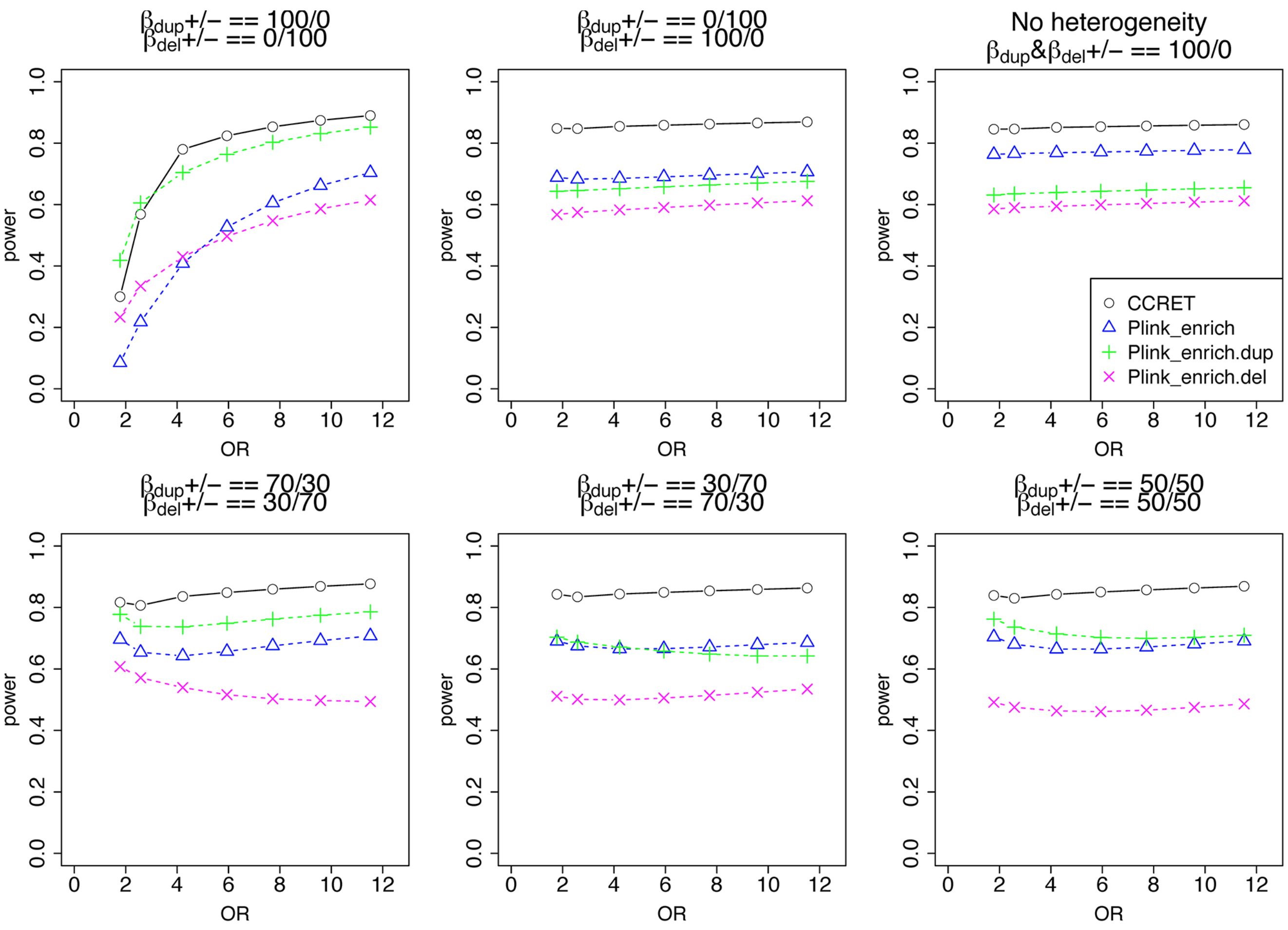


Fig 6. Power comparison between CCRET and PLINK 2-sided tests for simulation II-A: between-locus heterogeneity of the GI simulation, under 6 heterogeneity models as detailed in section "Simulation Design". Black **O** line: CCRET; Blue **Δ**: PLINK 2-sided test analyzing deletions and duplications combined; Green **+**: PLINK 2-sided test analyzing only duplications; Magenta **x**: PLINK 2-sided test analyzing only deletions.

doi:10.1371/journal.pgen.1005403.g006

of CCRET over PLINK.enrich, PLINK.enrich.dup and PLINK.enrich.del across all heterogeneity models. The consistent power gain in CCRET, which was not observed in the dosage simulation, was perhaps because the signal to noise ratio (SNR) in the GI simulation was smaller than that in the dosage simulation. Specifically, the SNR was 7.8% for this simulation but 34.1% in the dosage simulation. In SNP analysis, it has also been found that fixed effects approaches tend to be less powerful than random effects approaches when a high proportion of non-causal loci exist [19, 20]. Among the PLINK methods, PLINK-del has the lowest power in most of the scenarios, which is likely because in the causal genes, there are more duplication events intersecting the DupOnly causal genes than deletion events intersecting the DelOnly causal genes.

Simulation II-B: within-locus heterogeneity of the GI simulation. We used the 6 "Dup-Del" genes as causal and considered 3 heterogeneity models, i.e., (Dup, Del) = (P,R), (R,P) and (R,R) (no heterogeneity). We set a constant effect size for all casual loci and considered OR ranging between 3 and 50, chosen based on empirical evidences [4, 5]. The type I error rates (Table 2) were around the nominal level for PLINK.enrich methods and CCRET. The power results (Fig 7) suggest a consistent power gain of CCRET over PLINK.enrich based methods. The SNR was 1.2% for II-B and 2.2% in the I-B dosage simulation. In Fig 7, PLINK.del had the

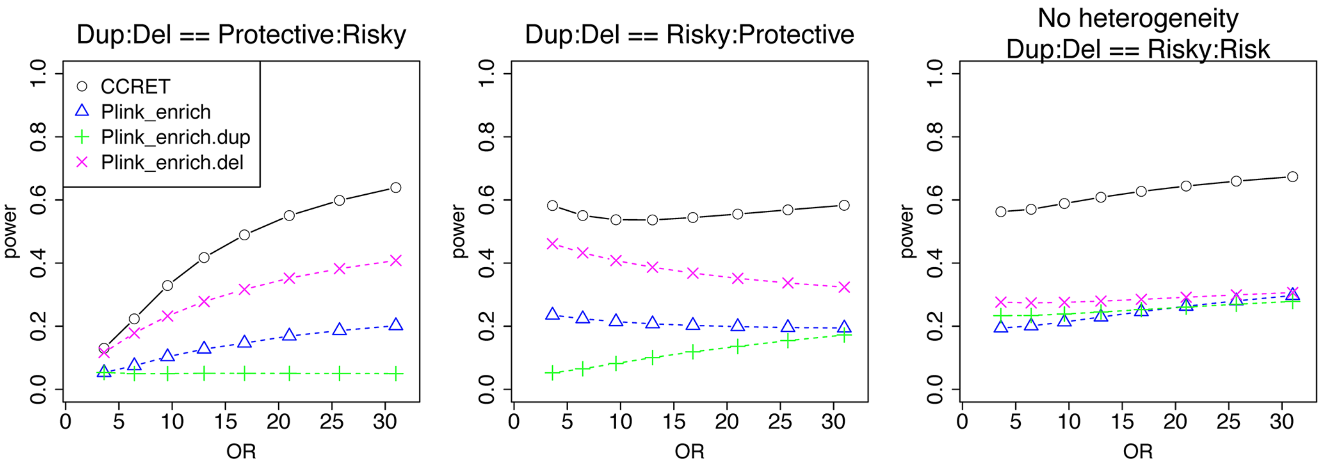


Fig 7. Power comparison between CCRET and PLINK 2-sided tests for simulation II-B: within-locus heterogeneity of the GI simulation, under 3 heterogeneity models as detailed in section "Simulation Design". Black **O** line: CCRET; Blue **Δ**: PLINK 2-sided test analyzing deletions and duplications combined; Green **+**: PLINK 2-sided test analyzing only duplications; Magenta **x**: PLINK 2-sided test analyzing only deletions.

doi:10.1371/journal.pgen.1005403.g007

highest power among PLINK tests since, among the 8 DupDel causal genes, there were more deletions than duplications. Note that we used the same aggregate functions as PLINK.enrich to model the background CNVs (i.e., dosage and size) in CCRET's GI model. Therefore the CCRET's GI model was the same as the PLINK.enrich model except the GI effect was modeled using a random effect. Consequently, the power gain in CCRET can be attributed to the use of random effects modeling of the GI effect.

Data application

We applied CCRET to CNV data from the Swedish schizophrenia study [36] and compared its performance to that of PLINK. The Swedish schizophrenia study used a case-control sampling design. Genotyping was done in six batches (Sw1-6) at the Broad Institute using Affymetrix 5.0 (3.9%, Sw1), Affymetrix 6.0 (38.6%, Sw2-4), and Illumina OmniExpress (57.4%, Sw5-6). We applied PennCNV [34] to generate CNV calls using the same protocol as we did for samples from the TwinGene project. After stringent quality control, we obtained a high-quality rare CNV (frequency <1%, size >100 Kb) dataset in 8,547 subjects (3,637 cases with schizophrenia and 4,820 controls) [36]. Further details on quality control can be found in ***S1 Text***. Previous analyses of this data indicated significant associations of large rare CNVs with schizophrenia risk for both genomewide dosage effects and GI effects of selected gene sets; and the associations were stronger when restricting to the largest categories (e.g. size > 500 kb) [9, 12]. CNVs with size >500 kb may be relatively more homogenous than CNVs with size > 100kb in the effects on disease risk as more causal genes may be impacted [9, 12].

We obtained the CNV data in PLINK format and converted it to the CCRET-format matrixes ("DS", "Len", "GI"). We verified that they stored identical information (i.e., the same total CNV numbers, average CNV lengths, and counts of intersected genes for each individual). We used the polynomial kernel with $d = 1$ and fit CCRET models (2) and (4) to evaluate the dosage effects and GI effects, respectively. The CCRET p-values were obtained using the Davies method [37]. We started by examining CNVs of size greater than 100kb, using CCRET and the 2-sided PLINK methods; and then repeated the analysis using CNVs of size greater than 500kb. In all analyses, we adjusted for the batch effect by including a batch indicator [9, 12]. As each genotyping batch used one specific type of arrays, controlling for genotyping batch effect simultaneously controlled for the difference in genotyping array. The comparison results are summarized in Table 3 for dosage effects and Table 4 for GI effects.

For dosage effects with CNVs that are >100kb, the results of PLINK.all and CCRET are different: PLINK.all (p-value 0.597) suggested no signal while CCRET had a significant p-value of 0.002. When focusing on CNVs that are >500kb, we replicated the literature finding of a significant association [9, 12] using both PLINK and CCRET: the p-value of PLINK.all is significant (p-value 0.013) while the p-value of CCRET is comparable (p-value 0.031). It seems that CNVs with size >500kb have relatively homogeneous dosage effects on the schizophrenia risk, where the case/control ratios of CNV rate are in the same directions for deletions-only and duplications-only (both >1.1, indicating an excess burden in cases). On the other hand, CNVs with size >100kb could have heterogeneous dosage effects, where the case/control ratios of CNV rate are in different directions for deletions-only (>1) and duplications-only (≤1). Consequently, PLINK.all had a significant result with >500kb CNVs but insignificant results with >100kb. In contrast, CCRET, due to its robustness with respect to the heterogeneity, yielded significant p-values for both types of CNVs. The results are not unexpected based on our findings in the simulations, where we observed that when CNVs have relatively homogeneous effects (which may occur in the case of CNVs >500kb in the real data), the best PLINK method can be more powerful than CCRET; when CNVs have heterogeneous effects (which may occur in the case of CNVs >100kb in the real data), CCRET is more powerful than PLINK.

For GI effects, we analyzed nine gene sets where significant enrichments of large rare CNVs in schizophrenia cases have been previously reported [12]. Further details on these gene sets can be found in ***S1 Text***. For most gene sets except for PSD-95, a similar pattern was observed with the GI analysis as with the dosage analysis: With CNVs > 500kb, we replicated the literature finding of significant associations [12] using both PLINK.enrich and CCRET. With CNVs > 100kb, we obtained different results between PLINK.enrich and CCRET,where CCRET often yielded significant p-values while PLINK.enrich did not. These results are not unexpected based on our findings in the simulations. For PSD-95, we observed significant results using PLINK. enrich for both CNVs > 100kb and CNVs > 500kb but insignificant results using CCRET. These results perhaps were influenced by the potentially high signal-to-noise-ratio in this small gene set (i.e. high proportion of causal loci) and relatively more homogenous CNV effects, which favors fixed effects collapsing methods over random effects methods.

Table 3. Test p-values for evaluating dosage effects based on schizophrenia data.

Size	Type	*#CNVs total*	*Case CNV rate*	*ControlCNV rate*	*Case/ controlratio*	*#CNVRstotal*	*#CNVRs with Dup. only*	*#CNVRs with Del. only*	*#CNVRs with DupDel*	*Pval_PLINK*	*Pval_CCRET*
>100Kb	DUP & DEL	8320	0.996	0.975	1.022	1853	1022 (55%)	517 (28%)	314 (17%)	0.597	0.002
	DUP	5070	0.599	0.600	0.997		1383			0.966	
	DEL	3250	0.397	0.375	1.061			897		0.192	
>500kb	DUP & DEL	840	0.110	0.091	1.199	300	202 (67%)	57 (19%)	41 (14%)	0.013	0.031
	DUP	617	0.080	0.068	1.175		246			0.062	
	DEL	223	0.030	0.024	1.267			99		0.085	

DUP: duplications. DEL: deletions. Case (or control) CNV rate = the total number of CNVs in cases (or controls) divided by the total number of cases (or controls). Case/control ratio = Case CNV rate divided by control CNV rate. CNVR: copy number variation region. Pval_PLINK: 2-sided p-values based on 10,000 permutations and permuting phenotype labels within genotyping batches (asymptotic p-values were similar). Pval_CCRET: 2-sided p-values based on the Davies (1980) method.

doi:10.1371/journal.pgen.1005403.t003

Table 4. Test p-values for evaluating GI effects based on schizophrenia data.

Gene-sets		CNVs	CNVs > 100 Kb						CNVs > 500 Kb					
Name (source)	#Genes	Type	#Genes any CNV	#GenesDup. only	#GenesDel. only	#GenesDupDel	Pval PLINK	Pval CCRET	#Genesany CNV	#Genes Dup.only	#GenesDel. only	#GenesDupDel	Pval PLINK	Pval CCRET
FRMP targets (Darnell et al.)	810	DUP & DEL	228	137	49	42	0.001	0.015	69	43	14	12	0.004	0.0006
		DUP		179			0.002			55			0.028	
		DEL			91		0.377				26		0.138	
PSD (Kirov et al)	668	DUP & DEL	153	100	27	26	0.110	0.018	55	29	14	12	0.002	0.008
		DUP		126			0.188			41			0.009	
		DEL			54		0.395				27		0.213	
PSD/mGluR5 (Kirov et al)	38	DUP & DEL	9	4	3	2	0.149	0.019	3	1		2	0.015	0.004
		DUP		6			0.225			3			0.048	
		DEL			5		0.622				2		0.295	
PSD/NMDAR (Kirov et al)	61	DUP & DEL	20	12	4	4	0.016	0.0007	9	4	3	2	0.002	0.002
		DUP		16			0.033			6			0.009	
		DEL			8		0.328				5		0.157	
PSD/PSD-95 (Kirov et al)	65	DUP & DEL	16	8	4	4	0.007	0.046	6	1	3	2	0.021	0.203
		DUP		12			0.344			3			0.727	
		DEL			8		0.008				5		0.001	
Mental Retardation	503	DUP & DEL	106	66	26	14	0.234	0.016	35	25	5	5	0.004	0.002
		DUP		80			0.099			3			0.004	
		DEL			40		0.661				1		0.750	
Synaptic genes (Ruano et al)	718	DUP & DEL	249	158	45	46	0.093	0.013	97	6	2	17	0.019	0.007
		DUP		203			0.457			77			0.337	
		DEL			92		0.060				38		0.002	
Synaptic Proteome (G2Cdb)	1023	DUP & DEL	177	112	38	27	0.0003	0.002	72	44	17	11	0.003	0.016
		DUP		139			0.002			55			0.025	
		DEL			66		0.086				29		0.117	
Cytoplasm (Kirov et al)	266	DUP & DEL	50	36	11	3	0.021	0.011	12	8	2	2	0.007	0.002
		DUP		39			0.008			1			0.006	
		DEL			14		0.821				4		0.574	

DUP: duplications. DEL: deletions. #Genes any CNV = the total number of genes intersected (≥1bp) by any CNV. #Genes Dup.only (or Del.only) = the total number of genes intersected (≥1bp) by duplications-only (or deletions-only). #Genes DupDel = the total number of genes intersected (≥ 1bp) by both duplications and deletions. Pval_PLINK: 2-sided asymptotic p-values based on Raychaudhuri et al (2010). Pval_CCRET: 2-sided p-values based on the Davies (1980) method.

doi:10.1371/journal.pgen.1005403.t004

Discussion

CNVs play an important role in disease etiology [4, 5]. While it is more informative to examine CNVs in the aggregate due to sparse events or modest marginal effects, current polygenic methods are mainly designed for bi-allelic SNPs and do not fully address the CNV-specific challenges. The challenges include the multiple features of a CNV affecting phenotypes, the non-biallelic nature of CNV polymorphisms, and etiological heterogeneity within and between loci. In this work, we introduce CCRET, a random effects method for CNV collapsing analysis that accommodates the poly-allelic nature of CNVs, models multiple CNV features and accounts for etiological heterogeneity. Simulation and real data analyses suggest that CCRET has stable performance across different scenarios of heterogeneity, and has a comparable or better power when compared to PLINK collapsing methods. The largest power gain tends to occur when heterogeneity pattern are complex, e.g., a mixture of risk-associated and protective effects observed within a locus or within a certain CNV type (duplication or deletion). The average running time for performing CCRET analysis with 4000 individuals is 27.5 seconds on an Intel Xeon 3.06 GHz machine with 64 Gb RAM.

With CCRET, we address the etiological heterogeneity by retaining the "locus"-specific effects and "allele"-specific effects during collapsing. To retain the locus-specific effect during collapsing, we define a locus unit for each CNV feature (e.g., CNVRs for dosage and genes for gene intersection); based on the locus definition, we quantify similarity shared between an individual pair for each locus and then aggregate the similarity information across loci. Because the multi-locus information is aggregated through a sum of genotype similarity instead of a sum of genotypes, loci with opposite effects do not cancel each other out. To retain allele-specific effect during collapsing, we factorize the allele values recorded in the design matrixes before calculating similarity. Factorizing the allele values avoids the need of dichotomizing the polymorphisms of a CNV feature and gains robustness against within-locus heterogeneity (as alleles with opposite effects do not need to be lumped together).

For GI analysis, we use genes instead of CNVRs as the locus unit. If CNVRs are used, the entry of the design matrix would record the number of genes intercepted by a certain CNV in the CNVR. Such data scoring may not be ideal when different intersected genes have different etiological effects. Hence we use genes as the locus unit so that the entry of the design matrix records the type of CNVs intersecting a gene. This allows us to model the gene-specific effects as well as the effects of the specific CNV type that intersects the gene.

Currently, there is no clear consensus on how to define CNVRs. One typical definition of CNVR is based on forming clusters of individual segments with some arbitrary amount of overlapping (e.g., 1 base pair overlap, 50% reciprocal overlap), and then allele frequency is computed for each CNVR. The choice of overlapping threshold could impact the formation of heterogeneous clusters that contain segments of variable sizes and breakpoints, each of which only overlap by a few base pairs but which may be disease relevant. For example, rare *NRXN1* deletions are associated with schizophrenia and show variable breakpoints and lengths among schizophrenia cases [38]. Such heterogeneous clusters of rare CNVs can occur frequently with high-resolution CNV detection technologies (e.g., high-density microarray and high-throughput sequencing); therefore defining CNVRs requires care [1, 4, 5]. In this work, we used the most inclusive threshold of 1 base-pair overlap to define CNVRs. Future study could account for CNVRs overlapping pattern by including them as fixed-effect covariates.

The simulation analyses show that, while sensitive to the underlying effect mechanism, the best PLINK (fixed effects) tests can be more powerful than CCRET when the effects of CNVs are homogeneous (such as the effects of >500kb CNVs in the real data analysis). The results agree with the findings in the SNP collapsing analysis; that is, the fixed effects approaches, which use the total event count to summarize the variant information within the target region, are the optimal methods when the majority of the variants in the region affect the phenotype in the same direction and with similar magnitude. On the other hand, when there exist non-causal variants (e.g., the effects of >100kb CNVs in the real data analysis) or a mixture of risk-associated and protective variants, random effects approaches are optimal because of their ability to account for effect heterogeneity. Taking together, we recommend that researchers apply both PLINK (fixed effects) and CCRET (random effects) in real-world rare CNV analysis because the underlying mechanisms of genetic effects are typically unknown.

Further, one future direction in which CCRET could be improved is to consider a hybrid approach, such as the SKAT-O test for rare SNP analysis [39]. Such a hybrid approach can retain the robustness of the random effects approach while retaining the test power of the fixed effects approach when the CNVs have homogeneous effects.

In the current work, we modeled the CNV feature of interest using random effects and the background CNV features using fixed effects. Alternatively, one can use random effects to model all CNV features; under this "fully random effects" model, the overall effect of each CNV feature is captured by a variance component. However, the calculation of the test statistics can be extremely computationally intensive, especially with large samples and non-normal traits, because it involves estimating several nuisance variance components. Similar computational concerns have also been encountered in GLMM based approaches for gene-environment interactions [40, 41]. We are exploring possible low-rank approximations to the CNV kernel matrixes using kernel principal component analysis [42] to reduce the computational burden in a fully random effects model.

To date, significant associations of rare CNVs with psychiatric disorders have been limited to the largest CNVs (e.g. >500 kb) [1, 4, 5]. In this work, the significant results with CNVs > 100kb obtained by CCRET are intriguing. There have been increasing evidences that smaller CNVs may contribute to the risk of psychiatric disorders [43, 44], although such associations have been challenging to detect because analytic methods are under developed. Improvement in CNV detection technologies will increase our ability to detect smaller CNVs (e.g., <20 kb), for which etiological heterogeneity can be more frequently encountered. The CCRET method could be important in analyzing smaller CNVs given its robustness in a variety of heterogeneous scenarios.

As we move ahead, success will increasingly depend on our ability to integrate all classes of genetic variation into a more complete disease model, including joint analysis of SNPs and CNVs [45]. For example, in schizophrenia genetics, multiple lines of genomic inquiry–genome-wide screens for rare CNVs, common SNPs, and rare exonic variation–are converging on similar sets of pathways and/or genes [12, 35, 36, 46]. The CCRET method may open a door for joint analysis as various variant types can be incorporated and simultaneously modeled under a random effects framework.

Materials and Methods

Ethics statement

All procedures were approved by ethical committees at the Karolinska Institutet in Sweden and at the University of North Carolina at Chapel Hill in the US, and all subjects provided written informed consent (or legal guardian consent and subject assent).

Input data format

As PLINK (version 1.07) is the standard software for CNV analysis from case/control data, we assume a PLINK-format CNV file as the source data, which lists the base pair position (start and end) and copy number (or dosage coded as 0,1,3,4+) of individual CNV segments. We create three input matrixes to store the different features of CNVs, i.e., dosage ("DS"), length ("Len"), and gene intersection ("GI"). For dosage and length matrixes, we first form the CNV region (CNVR) as the "locus" unit by clustering CNV segments using $\geq$1bp overlap from the PLINK format file. The (i, m)-entry of the DS matrix indicates the copy number (0, 1, 2, 3, 4+) of the CNV segment for subject i at CNVR m, and the (i, m)-entry of the Len matrix indicates the corresponding segment size in the DS matrix. The use of CNVRs allows us to keep track of CNVR-specific effects and to account for heterogeneity between and within CNVRs in the collapsing analysis.

For the gene intersection ("GI") matrix, instead of using CNVRs as we did for dosage and length, we use genes as the collapsing units to keep track of the effect of different genes when they are interrupted by CNVs. We obtained the coordinates of the genes of interest (http://genome.ucsc.edu/), and then create the GI matrix where the (i, m)-entry indicates whether gene m of subject i is intersected ($\geq$1bp overlapping) by a CNV and the corresponding CNV type (i.e., 0 for no intersection, 1 for intersection by a deletion, and 2 for intersection by a duplication). The genes of interest can be a collection of all protein coding genes or a collection of genes in a specific pathway. Using the gene as the locus unit allows us to evaluate the heterogeneous effects between and within genes in specific pathways. Alternatively, an exon could be used as the observation unit.

CCRET can handle CNVs called from both microarray and sequencing data. CNV files generated by CNV-calling algorithms from sequencing data are either in the Variant Call Format (VCF), such as those used by the 1000 Genomes project [3, 47], or in VCF-like but algorithm-specific output format (e.g., DATA.xcnv produced by the XHMM [48, 49] software from exome sequencing data). With CCRET, its data preparation pipeline provides scripts to convert any input CNV files first to PLINK-format CNV files as the source data and then to CCRET-specific CNV input matrixes (DS, Len, GI) for use in random effects modeling.

CCRET method

For subject i, let Y_i be a continuous or binary trait, X_i be a $p \times 1$ covariate vector including the intercept, and $Z_i^f = [Z_{i1}^f, \cdots, Z_{iM_f}^f]$ be a $M_f \times 1$ design vector of feature f in a certain genomic region, e.g., whole genome, pathway or certain type of genes. For $f = DS$ or Len, the design vector length, M_f, is the number of CNVRs in the targeted genomic region; for $f = GI$,M_{GI} is the number of genes in the targeted genomic region.

Suppose Y_i follows a distribution from the exponential family with density $f_Y(Y_i;\theta_i,\phi) = \exp[\{\theta_i Y_i - b(\theta\}/\{\phi v_i\}] + c(y_i,\phi)$, where θ_i is the canonical parameter with $\theta_i = g(\mu_i)$ with $g(\cdot)$ being a known link function, $\mu_i = E(Y_i \mid X_i, Z_i^{DS}, Z_i^{GI}, Z_i^{Len})$, $b(\cdot)$ and $c(\cdot)$ are known functions, ϕ is a dispersion parameter, and v_i is a known weight. The mean and variance of Y_i satisfy $\mu_i = b'(\theta_i)$ and $V(Y_i \mid X_i, Z_i^{DS}, Z_i^{GI}, Z_i^{Len}) = \phi v_i^{-1} b''(\theta_i)$ where $b'(\theta_i) = \frac{\partial}{\partial \theta_i} b(\theta_i)$ and $b''(\theta_i) = \frac{\partial}{\partial \theta_i} b'(\theta_i)$. We posit the following model to study the effects of CNV features on the trait values:

$$g(\mu_i) = \gamma_0 X_i + h_{DS}(Z_i^{DS}) + h_{GI}(Z_i^{GI}) + h_{Len}(Z_i^{Len}), \qquad (1)$$

where the covariate effects are modeled with effect size vector $\gamma_{0_{p\times1}}$, and $h_f(\cdot)$ with $f \in \{DS, GI, Len\}$ is a smooth function that models the effect of CNV feature f. There are many possible choices for the functions $h_f(\cdot)$. For example, one may set $h_f(Z_i^f) = \sum_{m=1}^{M_f} \beta_{f,m} Z_{im}^f$, allowing each CNVR to have its own effect $\beta_{f,m}$, therefore maximizing its ability to capture heterogeneity. This model corresponds to a classic linear regression but suffers from low power due to large degrees of freedom and sparse information. In order to reduce the degrees of freedom, one may impose a random effects model by assuming that the individual effects follow a normal distribution $\beta_{f,m} \sim N(0,\tau_f)$. Alternatively, Raychaudhuri et al.[27] (PLINK methods) considered the aggregated functions $h_{DS}(Z_i^{DS}) = \beta_{DS} \times \sum_{m=1}^{M_{DS}} I\{Z_{im}^{DS} \neq 2\}$ (where $\sum_{m=1}^{M_{DS}} I\{Z_{im}^{DS} \neq 2\}$ is the total number of CNVs of subject i) and $h_{GI}(Z_i^{GI}) = \beta_{GI} \times \sum_{m=1}^{M_{GI}} I\{Z_{im}^{GI} \neq 0\}$ (where $\sum_{m=1}^{M_{GI}} I\{Z_{im}^{GI} \neq 0\}$ is the number of genes intersected by CNVs for subject i), which nicely amplify the information content and avoid the problem of dimensionality by testing the averaged/common effect.

With CCRET, to avoid the dimensionality problem, we propose to model the covariates and background CNV features (measured from whole genome) using fixed effects, such as by setting $h_f(\cdot)$ as those considered in Raychaudhuri et al [27], but to model the CNV feature of interest (measured from the genomic regions of interest) using random effects. Specifically, when evaluating the effect of CNV feature f, we set $h_f(Z_i^f) \equiv h_i^f$, where h_i^f represents the subject-specific effect of multi-locus CNVs with feature f and is assumed to be random. Modeling the targeted CNV feature using random effects can capture the locus-specific details and account for between-locus and within-locus etiological heterogeneity when collapsing the information across different CNV regions. Treating the background CNV features as fixed effects can greatly boost the computational efficiency by bypassing the need to estimate the nuisance variance components, which can be burdensome with non-normal traits. As in the random effects approaches for SNP analyses (e.g., C-alpha [15], SKAT [16], SimReg [17, 18]), we assume that $h^f = (h_1^f, \cdots, h_n^f)^T \sim N(0, \tau_f K_f)$ where $K_f = \{ K_f(Z_i^f, Z_j^f) \}_{i,j=1}^n$ and $K_f(\cdot,\cdot)$ is a distance metric that quantifies the similarity between subject i and subject j based on CNV feature f in the targeted genomic region. In other words, the CNV information of feature f in the targeted region is first summarized by genetic similarity, which is then used to depict the covariance between CNV effects h_i^f and h_j^f. The aggregate CNV effect of feature f can be evaluated by examining the significance of the variance component τ_f (i.e., testing H_0: $\tau_f = 0$).

CCRET's connection with other random-effects collapsing methods

The proposed CCRET model has a direct connection with kernel machine regression [16, 50, 51] and gene-trait similarity regression [17, 18] because both kernel machine and similarity regressions have a mixed model representation. Specifically, under the kernel machine framework, our CCRET model is equivalent to specifying the CNV feature of interest, $h_f(\cdot)$, through a linear combination of kernel functions $K_f(\cdot,\cdot)$'s. That is, $h_f(Z_i^f) = \sum_{j=1}^n \alpha_j^f \times K_f(Z_i^f, Z_j^f)$ with α_j^f being the unknown parameter (the dual representation), or the equivalent basis representation, $h_f(Z_i^f) = \sum_{\ell=1}^L \eta_\ell^f \times \phi_\ell^f(Z_i^f)$, where $\{\phi_1^f(Z_i^f), \cdots, \phi_L^f(Z_i^f)\}$ is a set of the orthogonal basis functions spanning the functional space specified by $K_f(\cdot,\cdot)$ and η_ℓ^f is the unknown parameter. Under the similarity regression framework, where the genetic effect is assessed by the model significance that regresses trait similarity on genetic similarity, the variance component, τ_f, in our CCRET model is equivalent to the regression coefficient of genetic similarity quantified by the distance metric $K_f(\cdot,\cdot)$.

Quantification of CNV similarity between two individuals

From the connection with other random-effects collapsing methods, we see that one can determine how to model the multi-locus CNV information by selecting the desired metrics (kernels) to quantify similarity between subjects i and j. To fix the idea, consider the commonly used d-th order polynomial function, i.e., $K_f(Z_i^f, Z_j^f) = (1 + \sum_{m=1}^{M_f} w_m \times Z_{im}^f \times Z_{jm}^f)^d$, where w_m is the pre-specified weight for locus m based on, for example, some inverse function of the allele frequencies if CNVs of different frequency ranges are evaluated together. When $d = 1$, this corresponds to the model with only main effects, i.e., $h(Z_i^f) = \sum_{m=1}^{M_f} \eta_m^f \times Z_{im}^f$; when $d = 2$, this corresponds to a model with linear and quadratic main effects as well as two-way interactions, i.e., $h(Z_i^f) = \sum_{m=1}^{M_f} \eta_{1m}^f \times Z_{im}^f + \sum_{m=1}^{M_f} \eta_{2m}^f \times (Z_{im}^f)^2 + \sum_{\ell<m}^{M_f} \gamma_{\ell m}^f \times Z_{i\ell}^f \times Z_{im}^f$. For copy number dosage, it may not be sensible to use directly Z_{im}^{DS} and plug it in to the kernel function because both $Z_{im}^{DS} < 2$ and $Z_{im}^{DS} > 2$ deviate from normal copy number, while directly plugging Z_{im}^{DS} implies a dosage effect with copy number 0 as baseline. To resolve this issue, when quantifying similarity based on CNV dosage, we suggest to first covert Z_{im}^{DS} to a factorized design vector G_{im}^{DS}. For example, for a dosage ranging from 0 to 4, we define a length-3 design vector $G_{im}^{DS} = [1\ 0\ 0]$ for dosage < 2, $[0\ 1\ 0]$ for dosage $= 2$ and $[0\ 0\ 1]$ for dosage > 2; then $K_f(Z_i^{DS}, Z_j^{DS}) = (1 + \sum_{m=1}^{M_f} w_m \times G_{im}^{DS} T G_{jm}^{DS})^d$. The factorized design vector naturally accommodates the multi-categorical nature of dosage or GI. For discrete SNP data, one commonly used distance metric is the identity-by-state (IBS) metric, which is the proportion of alleles shared between two subjects in the targeted region. For CNV dosage, the IBS score at locus m becomes $G_{im}^T G_{jm}$ and is indeed the linear kernel.

Evaluating dosage effects

Below we illustrate the details of the proposed CCRET method using CNV dosage effects as an example, where we aim to assess the dosage effect while adjusting for the effects of CNV length and CNV gene counts. Following Raychaudhuri et al. [27], we define $\tilde{Z}_n^{GI}$ as the total number of genes that are intersected (i.e., including disrupted and overlapped) by the CNVs for subject i; define $\tilde{Z}_i^{Len}$ as the mean CNV length in kb of subject i. For those subjects with no CNVs (i.e., $\tilde{Z}_i^{Len} = 0$), their $\tilde{Z}_i^{Len}$ values are set to be the mean of the non-zero $\tilde{Z}_i^{Len}$' s. For dosage analysis, we rewrite Model (1) as

$$g(\mu_{1i}) = \gamma_1 X_i + \beta_{Len} \tilde{Z}_i^{Len} + \beta_{GI} \tilde{Z}_i^{GI} + h_i^{DS}, \tag{2}$$

where $h^{DS} = (h_1^{DS}, \cdots, h_n^{DS})^T \sim N(0, \tau_{DS} K_{DS})$ and K_{DS} is an $n \times n$ matrix with $K_{DS}(i, j) = K_{DS}(Z_i^{DS}, Z_j^{DS})$. Under Model (2), the dosage effect can be evaluated by testing $H_0^{DS} : \tau_{DS} = 0$. The incorporation of the background CNV features when assessing dosage effect is mainly for reducing the unexplained variance in the model and hence enhancing the detecting power. Using a similar derivation as Tzeng et al. [17], we construct a score-based test statistic to assess the dosage effect by considering a matrix presentation of Model (2):

$$g(\mu_1) = X\gamma_1 + \tilde{Z}^{Len} \beta_{Len} + \tilde{Z}^{GI} \beta_{GI} + h^{DS}, \tag{3}$$

where $\mu_1 = (\mu_{11}, \cdots, \mu_{1n})^T$, $X = (X_1, \cdots, X_n)^T$, $\tilde{Z}^{Len} = (\tilde{Z}_1^{len}, \cdots, \tilde{Z}_n^{len})^T$, and $\tilde{Z}^{GI} = (\tilde{Z}_1^{GI}, \cdots, \tilde{Z}_n^{GI})^T$. Using a very similar derivation of Tzeng and Zhang [52] and Tzeng et al. [17], it can be shown that the score test statistic is given as

$$T_{DS} = \frac{(Y - \mu_1)^T \Delta_1 W_1 K_{DS} W_1 \Delta_1 (Y - \mu_1))}{2} \Big|_{\tau_{DS}=0,\ \mu_1 = \widehat{\mu}_1,\ \phi_1 = \widehat{\phi}_1}$$

where $\mu_1 = g^{-1}(\tilde{X}_1 \theta_1)$ with $\tilde{X}_1 = (X, \tilde{Z}^{Len}, \tilde{Z}^{GI})$ and $\theta_1 = (\gamma_1, \beta_{Len}, \beta_{GI})^T$, $\Delta_1 = diag\{g'(\mu_{1i})\}$, and $W_1 = diag\{w_{1i}\}$, with $w_{1i} = \{\phi_1 m_{1i}^{-1} b''(\theta_{1i}))[g'(\mu_{1i})]^2\}^{-1}$. Estimate $\widehat{\theta}_1$ is the maximum likelihood estimate (MLE) of θ_1 under H_0, and $\widehat{\phi}_1$ is the restricted maximum likelihood (REML) type of estimate of ϕ_1 under H_0. As shown in Tzeng and Zhang [52], and Tzeng et al. [17], T_{DS}

asymptotically follows a weighted chi-squared distribution, i.e., $T_{DS} \approx \sum_{\ell=1}^{C} \lambda_{1\ell}\chi^2_{1,\ell}$, where $\lambda_{1\ell}$'s are the non-zero eigenvalues of $W_1^{-\frac{1}{2}} P_1 K_{DS} P_1 W_1^{-\frac{1}{2}}$ and $P_1 = W_1 - W_1 \tilde{X}_1 (\tilde{X}_1^T W_1 \tilde{X}_1)^{-1} \tilde{X}_1^T W_1$. The corresponding p-values can be obtained by Davies's methods (1980) [37] or by moment matching approaches as discussed in [53].

Evaluating gene-intersection (GI) effects

A very similar procedure can be used to assess the gene-intersection effect while adjusting for the effects of CNV length and the total number of CNV events. Specifically, we consider

$$g(\mu_{2i}) = \gamma_2 X_i + \delta_{Len} \tilde{Z}_i^{Len} + \delta_{DS} \tilde{Z}_i^{DS} + h_i^{GI}, \tag{4}$$

where $\tilde{Z}_i^{DS}$ is the total number of CNV events of subject i in the whole genome, $\tilde{Z}_i^{Len}$ is as defined before, $h^{GI} = (h_1^{GI}, \cdots, h_n^{GI})^T \sim N(0, \tau_{GI} K_{GI})$ and K_{GI} is an $n \times n$ matrix with $K_{GI}(i,j) = K_{GI}(Z_i^{GI}, Z_j^{GI})$. By a similar derivation as in the dosage analysis, the score statistics for testing = 0 can be obtained as:

$$T_{GI} = \frac{(Y-\mu_2)^T \Delta_2 W_2 K_{GI} W_2 \Delta_2 (Y-\mu_2)}{2} \Big|_{\tau_{GI}=0,\ \mu_2=\widehat{\mu}_2,\ \phi_2=\widehat{\phi}_2}$$

where $\mu_2 = g^{-1}(\tilde{X}_2 \theta_2)$ with $\tilde{X}_2 = (X, \tilde{Z}^{Len}, \tilde{Z}^{DS})$ and $\theta_2 = (\gamma_2, \delta_{Len}, \delta_{DS})^T$; Δ_2, W_2, $\widehat{\theta}_2$, and $\widehat{\phi}$ are defined in a fashion similar to the dosage test. The p-value of T_{GI} can also be obtained by Davies's methods (1980) [37] or by moment matching approaches as discussed in [53], because $T_{GI} \approx \sum_{\ell=1}^{C} \lambda_{2\ell}\chi^2_{1,\ell}$, where the $\lambda_{2\ell}$'s are the non-zero eigenvalues of $W_2^{-\frac{1}{2}} P_2 K_{GI} P_2 W_2^{-\frac{1}{2}}$ and $P_2 = W_2 - W_2 \tilde{X}_2 (\tilde{X}_2^T W_2 \tilde{X}_2)^{-1} \tilde{X}_2^T W_2$.

PLINK methods

We evaluated the performance of CCRET using the PLINK methods (version 1.07, http://pngu.mgh.harvard.edu/~purcell/plink/) as a benchmark. For dosage effect of the whole genome, we compared CCRET with the burden-style methods [9] as implemented in PLINK "—cnv-indiv-perm". This method fits a regression model: $(\mu_i) = \alpha_0 + \alpha_c \cdot c_i$, where c_i is the total number of events that are of interest (e.g., deletion, duplication, duplication+deletion etc.) for individual i. The "—cnv-indiv-perm" evaluates the significance of α_c via a permutation procedure. The default option returns 1-sided empirical p-values, assuming that the events of interest are more common in cases than in controls (i.e. events increase risk). The default 1-sided tests have been commonly adopted in practice [8–12]. Adding the flag "—cnv-test-2sided" will return 2-sided empirical p-values, allowing that the events of interest might be more common either in cases or in controls. For the GI effect of a gene set, we compared CCRET with the enrichment-style method of Raychaudhuri et al [27], which is implemented as "—cnv-enrichment-test" test in PLINK. This method fits a logistic regression model: $g(\mu_i) = \beta_0 + \beta_{DS} \cdot \tilde{Z}_i^{DS} + \beta_{Len} \cdot \tilde{Z}_i^{Len} + \beta_{GI} \cdot g_i$, where $\tilde{Z}_i^{DS}$ and $\tilde{Z}_i^{Len}$ are as defined earlier, g_i is the total number of intersected genes in a predefined gene set and β_f's are regression coefficients. The "—cnv-enrichment-test" tests if β_{GI}, the coefficient associated with GI counts, is significantly different from 0. The default option returns 2-sided asymptotic p-values, allowing that gene intersection might be more common either in cases or in controls.

Implementation

Source code of CCRET is available at http://www4.stat.ncsu.edu/~tzeng/software.php.

Supporting Information

S1 Text. This section includes the detailed information of the datasets used in this study. (DOCX)

Acknowledgments

We thank anonymous reviewers for their helpful comments. We also thank the TwinGene study for the use of CNV data in our study.

The Swedish Schizophrenia Study consortium: Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L. Moran, Anna K. Kähler, Susanne Akterin, Sarah Bergen, Patrik K.E. Magnusson, Benjamin M. Neale, Douglas Ruderfer, Menachem Fromer, Edward Scolnick, Shaun Purcell, Steve McCarroll, Pamela Sklar, Christina M. Hultman, and Patrick F. Sullivan.

Author Contributions

Conceived and designed the experiments: JYT JPS. Performed the experiments: JPS JYT. Analyzed the data: JPS JYT. Contributed reagents/materials/analysis tools: JYT JPS PKEM PFS. Wrote the paper: JYT JPS PKEM PFS.

References

1. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011; 12(5):363–76. Epub 2011/03/02. doi: 10.1038/nrg2958 PMID: 21358748.
2. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470(7332):59–65. Epub 2011/02/05. doi: nature09708 [pii] doi: 10.1038/nature09708 PMID: 21293372; PubMed Central PMCID: PMC3077050.
3. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. doi: 10.1038/nature11632 PMID: 23128226; PubMed Central PMCID: PMC3498066.
4. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nature Reviews Genetics. 2012; 13:537–51. PMID: 22777127. doi: 10.1038/nrg3240
5. Malhotra D, Sebat J. CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. Cell. 2012; 148(6):1223–41. Epub 2012/03/20. doi: 10.1016/j.cell.2012.02.039 PMID: 22424231.
6. Glessner JT, Connolly JJ, Hakonarson H. Rare genomic deletions and duplications and their role in neurodevelopmental disorders. Curr Top Behav Neurosci. 2012; 12:345–60. doi: 10.1007/7854_2011_179 PMID: 22241247.
7. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010; 11(11):773–85. doi: 10.1038/nrg2867 PMID: 20940738; PubMed Central PMCID: PMC3743540.
8. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008; 320:539–43. PMID: 18369103. doi: 10.1126/science.1155174
9. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008; 455:237–41. PMID: 18668038. doi: 10.1038/nature07239
10. Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P, et al. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum Mol Genet. 2009; 18(8):1497–503. PMID: 19181681. doi: 10.1093/hmg/ddp043
11. Buizer-Voskamp JE, Muntjewerff JW, Genetic R, Outcome in Psychosis Consortium M, Strengman E, Sabatti C, et al. Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. Biol Psychiatry. 2011; 70(7):655–62. doi: 10.1016/j.biopsych.2011.02.015 PMID: 21489405; PubMed Central PMCID: PMC3137747.
12. Szatkiewicz JP, O'Dushlaine C, Chen G, Chambert K, Moran JL, Neale BM, et al. Copy number variation in schizophrenia in Sweden. Mol Psychiatry. 2014. doi: 10.1038/mp.2014.40 PMID: 24776740.

13. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–21. doi: 10.1016/j.ajhg.2008.06.024 PMID: 18691683; PubMed Central PMCID: PMC2842185.

14. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86(6):832–8. doi: 10.1016/j.ajhg.2010.04.005 PMID: 20471002; PubMed Central PMCID: PMC3032073.

15. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genetics. 2011;e1001322. doi: 10.1371/journal.pgen.1001322 PMID: 21408211

16. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. doi: 10.1016/j.ajhg.2011.05.029 PMID: 21737059; PubMed Central PMCID: PMC3135811.

17. Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65(3):822–32. doi: 10.1111/j.1541-0420.2008.01176.x PMID: 19210740; PubMed Central PMCID: PMC2748404.

18. Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. American journal of human genetics. 2011; 89(2):277–88. Epub 2011/08/13. doi: 10.1016/j.ajhg.2011.07.007 PMID: 21835306; PubMed Central PMCID: PMC3155192.

19. Pongpanich M, Neely ML, Tzeng JY. On the Aggregation of Multimarker Information for Marker-Set and Sequencing Data Analysis: Genotype Collapsing vs. Similarity Collapsing. Front Genet. 2011; 2:110. doi: 10.3389/fgene.2011.00110 PMID: 22303404; PubMed Central PMCID: PMC3266618.

20. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014; 95(1):5–23. doi: 10.1016/j.ajhg.2014.06.009 PMID: 24995866; PubMed Central PMCID: PMC4085641.

21. Girirajan S, Johnson RL, Tassone F, Balciuniene J, Katiyar N, Fox K, et al. Global increases in both common and rare copy number load associated with autism. Hum Mol Genet. 2013; 22(14):2870–80. doi: 10.1093/hmg/ddt136 PMID: 23535821; PubMed Central PMCID: PMC3690969.

22. Bassett AS, Chow EW, Husted J, Weksberg R, Caluseriu O, Webb GD, et al. Clinical features of 78 adults with 22q11 Deletion Syndrome. Am J Med Genet A. 2005; 138(4):307–13. doi: 10.1002/ajmg.a.30984 PMID: 16208694; PubMed Central PMCID: PMC3127862.

23. Murphy KC, Jones RG, Griffiths E, Thompson PW, Owen MJ. Chromosome 22qII deletions. An underrecognised cause of idiopathic learning disability. Br J Psychiatry. 1998; 172:180–3. PMID: 9519073.

24. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J, et al. Copy number variants in schizophrenia: Confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am J Psychiatry. 2011; 168:302–16. Epub 2011/02/03. appi.ajp.2010.10060876 [pii] doi: 10.1176/appi.ajp.2010.10060876 PMID: 21285140.

25. Rees E, Kirov G, Sanders A, Walters JT, Chambert KD, Shi J, et al. Evidence that duplications of 22q11.2 protect against schizophrenia. Mol Psychiatry. 2014; 19(1):37–40. doi: 10.1038/mp.2013.156 PMID: 24217254; PubMed Central PMCID: PMC3873028.

26. Vacic V, McCarthy S, Malhotra D, Murray F, Chou HH, Peoples A, et al. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature. 2011; 471(7339):499–503. Epub 2011/02/25. doi: 10.1038/nature09884 PMID: 21346763.

27. Raychaudhuri S, Korn JM, McCarroll SA, International Schizophrenia C, Altshuler D, Sklar P, et al. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet. 2010; 6(9):e1001097. doi: 10.1371/journal.pgen.1001097 PMID: 20838587; PubMed Central PMCID: PMC2936523.

28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007; 81:559–75. PMID: 17701901

29. Beekman M, Heijmans BT, Martin NG, Whitfield JB, Pedersen NL, DeFaire U, et al. Two-locus linkage analysis applied to putative quantitative trait loci for lipoprotein(a) levels. Twin Res. 2003; 6(4):322–4. doi: 10.1375/136905203322296692 PMID: 14511440.

30. Heijmans BT, Beekman M, Putter H, Lakenberg N, van der Wijk HJ, Whitfield JB, et al. Meta-analysis of four new genome scans for lipid parameters and analysis of positional candidates in positive linkage regions. Eur J Hum Genet. 2005; 13(10):1143–53. doi: 10.1038/sj.ejhg.5201466 PMID: 16015283.

31. Lichtenstein P, Bjork C, Hultman CM, Scolnick EM, Sklar P, Sullivan PF. Recurrence risks for schizophrenia in a Swedish national cohort. Psychol Med. 2006; 36:1417–26. PMID: 16863597.

32. Lichtenstein P, Sullivan P, Cnattingius S, Gatz M, Johansson S, Carlström C, et al. The Swedish Twin Registry in the Third Millennium–an update. Twin Res Hum Genet. 2006; 9:875–82. PMID: 17254424

33. Pedersen NL, Lichtenstein P, Svedberg P. The Swedish Twin Registry in the Third Millenium. Twin Research. 2002; 5:427–32. PMID: 12537870

34. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007; 17(11):1665–74. PMID: 17921354.

35. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Molecular psychiatry. 2011. Epub 2011/11/16. doi: 10.1038/mp.2011.154 PMID: 22083728.

36. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013; 45(10):1150–9. doi: 10.1038/ng.2742 PMID: 23974872; PubMed Central PMCID: PMC3827979.

37. Davies RB. Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables,. Journal of the Royal Statistical Society Series C (Applied Statistics). 1980; 29(3):323–33.

38. Kirov G, Rujescu D, Ingason A, Collier DA, O'Donovan MC, Owen MJ. Neurexin 1 (NRXN1) deletions in schizophrenia. Schizophr Bull. 2009; 35(5):851–4. Epub 2009/08/14. doi: 10.1093/schbul/sbp079 PMID: 19675094; PubMed Central PMCID: PMC2728827.

39. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13(4):762–75. doi: 10.1093/biostatistics/kxs014 PMID: 22699862; PubMed Central PMCID: PMC3440237.

40. Zhao G, Marceau R, Zhang D, Tzeng JY. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. Genetics. 2015; 199(3):695–710. doi: 10.1534/genetics.114.171686 PMID: 25585620; PubMed Central PMCID: PMC4349065.

41. Lin X, Lee S, Christiani DC, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics. 2013; 14(4):667–81. doi: 10.1093/biostatistics/kxt006 PMID: 23462021; PubMed Central PMCID: PMC3769996.

42. Bernhard Schölkopf AS, Er Smola, Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation. 1998; 10:1299–319.

43. Poultney CS, Goldberg AP, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwara Y, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am J Hum Genet. 2013; 93(4):607–19. doi: 10.1016/j.ajhg.2013.09.001 PMID: 24094742; PubMed Central PMCID: PMC3791269.

44. Szatkiewicz JP, Neale BM, O'Dushlaine C, Fromer M, Goldstein JI, Moran JL, et al. Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. Mol Psychiatry. 2013; 18(11):1178–84. doi: 10.1038/mp.2013.98 PMID: 23938935; PubMed Central PMCID: PMC3966073.

45. Gamazon ER, Cox NJ, Davis LK. Structural architecture of SNP effects on complex traits. Am J Hum Genet. 2014; 95(5):477–89. doi: 10.1016/j.ajhg.2014.09.009 PMID: 25307299; PubMed Central PMCID: PMC4225594.

46. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014; 506(7487):185–90. doi: 10.1038/nature12975 PMID: 24463508.

47. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. Epub 2010/10/29. doi: 10.1038/nature09534 PMID: 20981092; PubMed Central PMCID: PMC3042601.

48. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012; 91(4):597–607. Epub 2012/10/09. doi: 10.1016/j.ajhg.2012.08.005 PMID: 23040492; PubMed Central PMCID: PMC3484655.

49. Fromer M, Purcell SM. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr Protoc Hum Genet. 2014; 81:7 23 1–7 1. doi: 10.1002/0471142905.hg0723s81 PMID: 24763994; PubMed Central PMCID: PMC4065038.

50. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63(4):1079–88. doi: 10.1111/j.1541-0420.2007.00799.x PMID: 18078480; PubMed Central PMCID: PMC2665800.

51. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008; 9:292. doi: 10.1186/1471-2105-9-292 PMID: 18577223; PubMed Central PMCID: PMC2483287.

52. Tzeng JY, Zhang D. Haplotype-based association analysis via variance-components score test. Am J Hum Genet. 2007; 81(5):927–38. doi: 10.1086/521558 PMID: 17924336; PubMed Central PMCID: PMC2265651.

53. Pierre Duchesne PLDM. Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. Computational Statistics and Data Analysis. 2010; 54(4):858–62.